Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.00093

TƯ VẤN LỌC CỘNG TÁC DỰA TRÊN NGƯỜI SỬ DỤNG DÙNG PHÉP ĐO GẮN KẾT HÀM Ý THỐNG KÊ

Phan Phương Lan[1], Trần Uyên Trang[2], Huỳnh Hữu Hưng[3,4], Huỳnh Xuân Hiệp[1]

[1]Trường Đại học Cần Thơ, Việt Nam
[2]Trường Đại học Sư phạm Đà Nẵng, Đại học Đà Nẵng, Việt Nam
[3]Trường Đại học Bách khoa Đà Nẵng, Đại học Đà Nẵng, Việt Nam
[4]Viện Nghiên cứu và Đào tạo Việt - Anh, Đại học Đà Nẵng, số 41 Lê Duẩn, Hải Châu, Đà Nẵng, Việt Nam
pplan@cit.ctu.edu.vn, trang.tranuyen@gmail.com, hhhung@dut.udn.vn, hxhiep@ctu.edu.vn

TÓM TẮT—*Phép đo tương tự giữ vai trò chính trong hệ tư vấn lọc cộng tác. Hệ số tương quan Pearson, độ tương tự Cosine và chỉ số Jacard là những phép đo tương tự phổ biến được dùng để tính sự tương đồng giữa hai người sử dụng. Những phép đo này đều có đặc điểm đối xứng trong khi thực tế, sự ảnh hưởng qua lại giữa hai người sử dụng thường không đối xứng. Bài báo đề xuất một cách tiếp cận mới trong phương pháp lọc cộng tác để tính sự tương tự cho từng cặp người sử dụng thông qua phép đo gắn kết hàm ý thống kê Cohesion, đồng thời tập trung vào việc đánh giá mô hình hệ tư vấn dùng phép đo được đề xuất với các mô hình hệ tư vấn dùng những phép đo tương tự Cosine, Pearson, và Jaccard trên tập liệu 0 – 1. MSWeb được chọn làm tập dữ liệu thực nghiệm và k-fold cross validation được sử dụng làm phương pháp phân tách dữ liệu. Kết quả thực nghiệm cho thấy mô hình hệ tư vấn lọc cộng tác dựa trên người sử dụng dùng phép đo Cohesion có ưu thế hơn so với các mô hình hệ tư vấn dùng các phép đo còn lại.*

Từ khóa— *Lọc cộng tác dựa trên người sử dụng, phép đo gắn kết hàm ý thống kê, hệ tư vấn.*

I. GIỚI THIỆU

Hệ tư vấn là kỹ thuật và công cụ phần mềm đề xuất những mục dữ liệu mà người sử dụng có thể muốn [8]. Hiện nay, hệ tư vấn được sử dụng rộng rãi trong nhiều dịch vụ, chẳng hạn như eBay hay Amazon. Với một tập người sử dụng (user), một tập các mục dữ liệu (item), và các đánh giá (rating) tường minh hay không tường minh nhằm thể hiện mức độ người sử dụng thích hay không thích một mục dữ liệu mà người đó đã xem (hoặc nghe hoặc mua, v.v.), hệ tư vấn dự đoán mức đánh giá cho một mục chưa được xem bởi người sử dụng, hoặc cung cấp một danh sách các mục mà người sử dụng có thể thích. Các phương pháp tư vấn được phân thành những nhóm chính: tư vấn dựa trên nội dung, tư vấn lọc cộng tác và tư vấn dạng hỗn hợp [2] [8] [9] mà trong đó lọc cộng tác [2] [8] [9] [13] là phương pháp quan trọng và được sử dụng phổ biến nhất. Một trong những cách tiếp cận chính của lọc cộng tác là lọc cộng tác dựa trên người sử dụng [7]. Cách tiếp cận này sử dụng trực tiếp những đánh giá được lưu trữ trong hệ thống; và đề cử một danh sách các mục dữ liệu cho một người sử dụng hoặc dự đoán mức đánh giá cho một mục cụ thể dựa trên phép đo tương tự giữa những người sử dụng. Như vậy, các phép đo tương tự giữ vai trò chính trong các hệ tư vấn lọc cộng tác. Hệ số tương quan Pearson, độ tương tự Cosine và chỉ số Jacard [2] [6] [8] là các phép đo tương tự phổ biến. Tất cả những phép đo này đều có đặc điểm đối xứng. Điều này có nghĩa là, với một cặp người sử dụng, sự ảnh hưởng của một người lên người còn lại là như nhau. Tuy nhiên, thực tế thì không như vậy, sự ảnh hưởng qua lại giữa hai người thường không đối xứng. Người sử dụng này có thể đánh giá các mục dữ liệu khá giống với người sử dụng kia nhưng điều ngược lại có thể không đúng.

Phép đo gắn kết (Cohesion) được đề xuất đầu tiên trong phân tích hàm ý thống kê [19]. Mục đích của phép đo Cohesion là phát hiện ra những luật kết hợp có chất lượng hàm ý tốt, nói cách khác là phát hiện ra các luật $A \rightarrow B$ có mối quan hệ hàm ý mạnh (gắn kết mạnh) giữa A và B. Cohesion là một phép đo không đối xứng, giá trị gắn kết của (A, B) và của (B, A) là độc lập. Do những đặc trưng vừa nêu, phép đo Cohesion có thể vận dụng vào việc xác định mức độ tương tự (gắn kết) giữa hai người sử dụng trong lọc cộng tác.

Bài báo này đề xuất một cách tiếp cận mới để tính độ tương tự cho từng cặp người sử dụng dùng trong phương pháp lọc cộng tác, đồng thời tập trung vào việc đánh giá các mô hình hệ tư vấn dùng các phép đo khác nhau.

Bài báo được tổ chức thành 5 phần. Phần I giới thiệu chung. Phần II mô tả phương pháp tư vấn lọc cộng tác dựa trên người sử dụng dùng phép đo gắn kết hàm ý thống kê. Phần III tập trung vào đánh giá hệ tư vấn lọc cộng tác dựa trên người sử dụng. Phần IV là thực nghiệm để đánh giá các mô hình hệ tư vấn lọc cộng tác dùng các phép đo Cosine, Pearson, Jaccard, và Cohesion để xác định sự tương tự giữa ha người sử dụng. Phần cuối cùng đưa ra kết luận.

II. TƯ VẤN LỌC CỘNG TÁC DỰA TRÊN NGƯỜI SỬ DỤNG DÙNG PHÉP ĐO GẮN KẾT HÀM Ý THỐNG KÊ

A. Mô hình hệ tư vấn

Mô hình hệ tư vấn gồm các thành phần: tập người sử dụng $U = \{u_1, u_2, \ldots, u_m\}$, tập các mục dữ liệu $I = \{i_1, i_2, \ldots, i_n\}$, và ma trận đánh giá $R = (r_{jk})$ với $j = 1..m$ và $k = 1..n$ lưu kết quả đánh giá của người sử dụng u_j cho một mục dữ liệu i_k. Trong ma trận R, r_{jk} có thể rỗng nếu người sử dụng u_j chưa đánh giá mục dữ liệu i_k. Việc người sử dụng không muốn tiết lộ trực tiếp những sở thích của họ thông qua đánh giá mục dữ liệu là tình huống chung.

Trong trường hợp như thế, các sở thích chỉ có thể được suy luận bằng cách phân tích hành vi của người sử dụng, chẳng hạn như: cần/không cần (biết/không biết) một mặt hàng, nhấp chuột/không nhấp chuột vào một mục nào đó của website. Các hành vi này sẽ tạo ra các dữ liệu 0 – 1.

Với các tập dữ liệu 0 – 1, ma trận đánh giá R được biểu diễn như Bảng 1. $R = (r_{jk})$ với $r_{jk} = \{0 \text{ hoặc } 1\}$, $j = 1..m$ và $k = 1..n$. Giá trị 0 có nghĩa là mục dữ liệu không được thích (không được cần, không được biết) bởi người sử dụng; và giá trị 1 có nghĩa là mục dữ liệu được thích bởi người sử dụng.

Bảng 1. Ma trận đánh giá lưu kết quả đánh giá của người sử dụng cho các mục dữ liệu

	$\mathbf{i_1}$	$\mathbf{i_2}$	$\mathbf{i_3}$		$\mathbf{i_n}$
$\mathbf{u_1}$	**0**	**1**	**0**	…	**1**
$\mathbf{u_2}$	**1**	**0**	**1**	…	**0**
…	…	…	…	…	…
$\mathbf{u_m}$	**1**	**0**	**0**	…	**1**

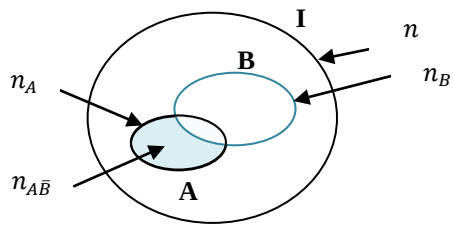
B. Phép đo gắn kết hàm ý thống kê

Phép đo gắn kết hàm ý thống kê Cohesion [19] – một phép đo không đối xứng – được sử dụng để phát hiện các luật $A \rightarrow B$ có mối quan hệ hàm ý mạnh giữa A và B. Bài báo này đề xuất việc sử dụng phép đo gắn kết hàm ý thống kê vào tư vấn lọc cộng tác dựa trên người sử dụng.

1. Biểu diễn mối quan hệ hàm ý giữa hai người sử dụng

Đặt $A \subset I$ là tập gồm những mục dữ liệu được đánh giá bởi người sử dụng u_a; $\overline{A}$ là tập bù của A; $B \subset I$ là tập gồm những mục dữ liệu được đánh giá bởi người sử dụng u_b; $\overline{B}$ là tập bù của B; $n_A = \text{card}(A)$ là số mục dữ liệu được đánh giá bởi người sử dụng u_a (số phần tử của A); $n_B = \text{card}(B)$ là số mục dữ liệu được đánh giá bởi người sử dụng u_b (số phần tử của B); và $n_{A\overline{B}} = \text{card}(A \cap \overline{B})$ là số mục dữ liệu được đánh giá bởi người sử dụng u_a mà không (chưa) được đánh giá bởi người sử dụng u_b (số phản ví dụ).

Mối quan hệ hàm ý giữa người sử dụng u_a với u_b, nói cách khác là mối quan hệ giữa tập mục A được thích bởi người sử dụng u_a với tập mục B được thích bởi người sử dụng u_b, được biểu diễn bởi một bộ gồm bốn phần tử $\{n, n_A, n_B, n_{A\overline{B}}\}$.



Hình 1. Các phần tử biểu diễn mối quan hệ hàm ý giữa người sử dụng u_a với u_b (mối quan hệ $A \rightarrow B$)

2. Đo sự gắn kết hàm ý giữa hai người sử dụng

Để đo sự gắn kết hàm ý thống kê giữa người sử dụng u_a với u_b, phép đo gắn kết Cohesion được sử dụng. Phép đo này được ký hiệu là $c(u_a, u_b)$, và có công thức tính như (1).

$$c(u_a, u_b) = \begin{cases} (1 - (-p\log_2 p - (1-p)\log_2(1-p))^2)^{1/2} \text{ nếu } p > 0.5 \\ 0 \text{ nếu ngược lại} \end{cases} \quad (1)$$

Trong đó, $p = \varphi(u_a, u_b)$ là mật độ hàm ý (implicative intensity).

Phép đo mật độ hàm ý $\varphi(u_a, u_b)$ được sử dụng để đo khuynh hướng các mục dữ liệu được đánh giá bởi người sử dụng u_b khi chúng được đánh giá bởi người sử dụng u_a. $\varphi(u_a, u_b)$ được tính theo công thức (2) với $q(A, \overline{B})$ và λ được xác định theo công thức (3) và (4) tương ứng.

$$\varphi(u_a, u_b) = \begin{cases} 1 - \sum_{s=0}^{n_{A\overline{B}}} \frac{\lambda^s}{s!} e^{-\lambda} = \frac{1}{2\pi} \int_{q(A,\overline{B})}^{\infty} e^{-\frac{t^2}{2}} dt \text{ nếu } n_B \neq n \\ 0 \text{ nếu ngược lại} \end{cases} \quad (2)$$

$$q(A,\overline{B}) = \frac{n_{A\overline{B}} - \frac{n_A(n-n_B)}{n}}{\sqrt{\frac{n_A(n-n_B)}{n}}} \quad (3) \qquad \lambda = \frac{n_A(n-n_B)}{n} \quad (4)$$

C. Tư vấn lọc cộng tác sử dụng phép đo gắn kết hàm ý thống kê

Với phương pháp lọc cộng tác dựa trên người sử dụng, để đề xuất các mục dữ liệu cho một người sử dụng cụ thể u, một tập những người sử dụng khác tương tự với u sẽ được tham khảo. Các mục dữ liệu được đề xuất cho u là những mục mà u có thể thích nhưng đã không nhìn thấy hoặc không biết. Giải thuật xây dựng hệ tư vấn lọc cộng tác dựa trên người sử dụng dùng phép đo gắn kết hàm ý thống kê được đề xuất như sau:

Input: ma trận đánh giá R bao gồm thông tin đánh giá của người sử dụng u (các mục dữ liệu mà u đã đánh giá)

Output: danh sách các mục được đề xuất cho người sử dụng u.

Các bước thực hiện:

- Tính sự tương tự giữa người sử dụng u với tất cả những người sử dụng khác bằng cách sử dụng phép đo gắn kết Cohesion.
- Chọn ra k người sử dụng tương tự nhất với người sử dụng u (k láng giềng gần nhất).
- Dự đoán mức đánh giá của người sử dụng u cho các mục dữ liệu chưa được đánh giá bằng cách sử dụng các đánh giá của k người dùng tương tự nhất.
- Đề xuất một danh sách các mục dữ liệu cho người sử dụng u dựa trên việc sắp xếp các đánh giá được dự đoán.

Tại bước 1 của giải thuật trên, phép đo gắn kết Cohesion được sử dụng để tính sự tương tự giữa hai người sử dụng. Bài báo này hướng tới vào việc sử dụng các tập dữ liệu 0 – 1 nên giải thuật tính giá trị của phép đo gắn kết giữa người sử dụng u_a với u_b được đề xuất như sau:

Input: thông tin đánh giá của người sử dụng u_a và u_b.

Output: giá trị gắn kết giữa người sử dụng u_a với u_b.

Các bước thực hiện:

- Xác định giá trị của các phần tử $n, n_A, n_B, n_{A\overline{B}}$ biểu diễn cho mối quan hệ hàm ý giữa người sử dụng u_a với u_b

  n = số cột của ma trận R $n_A = \sum_{k=1}^{n} R[u_a][k]$;

  $n_B = \sum_{k=1}^{n} R[u_b][k]$; $n_{A\overline{B}} = n_A - \sum_{k=1}^{n} R[u_a][k]R[u_b][k]$;
- Tính mật độ hàm ý của mối quan hệ giữa người sử dụng u_a với u_b,

  $\lambda = (n_A(n-n_B))/n$; $sum = \sum_{s=0}^{n_{A\overline{B}}} \frac{\lambda^s}{s!} e^{-\lambda}$; $p = 1 - sum$
- Đo sự gắn kết giữa người sử dụng u_a với u_b

  if (p<=0.5) return 0;

  else return $(1-(-p\log_2 p-(1-p)\log_2(1-p))^2)^{1/2}$

III. ĐÁNH GIÁ HỆ TƯ VẤN LỌC CỘNG TÁC DỰA TRÊN NGƯỜI SỬ DỤNG

A. Phép đo dùng để đánh giá hệ tư vấn

Có nhiều cách để đánh giá hệ tư vấn, trong đó đánh giá offline thường giữ vai trò chính [6] [10]. Đánh giá offline sử dụng tập dữ liệu đã được thu thập trước, không cần sự tương tác với người sử dụng thật nên chi phí cho đánh giá hệ tư vấn là thấp.

Các hệ tư vấn dựa trên dữ liệu 0 -1 thường tập trung vào việc đề xuất cho người sử dụng những mục dữ liệu mà họ có thể thích hơn là dự đoán sở thích của người sử dụng cho một mục dữ liệu cụ thể. Các phép đo dùng để đánh giá tính chính xác của việc đề xuất là độ bao phủ (recall) và độ chính xác (precision). Những phép đo này được xây dựng dựa vào ma trận nhầm lẫn như Bảng 2. Các thành phần của ma trận nhầm lẫn là (1) - TN: số mục dữ liệu không được hệ tư vấn đề xuất và cũng không được người sử dụng thích; (2) - FP: số mục dữ liệu được hệ tư vấn đề xuất nhưng không được người sử dụng thích; (3) - FN: số mục dữ liệu không được hệ tư vấn đề xuất nhưng được người sử dụng thích; (4) - TP: số mục dữ liệu được hệ tư vấn đề xuất và cũng được người sử dụng thích.

Bảng 2. Ma trận nhầm lẫn

Thực tế / Dự đoán	Không được đề xuất	Được đề xuất
Không được thích	**TN**	**FP**
Được thích	**FN**	**TP**

Công thức của để tính giá trị của phép đo bao phủ và phép đo chính xác được trình bày trong (6) và (7) tương ứng.

$$\text{Độ chính xác } = \frac{\{\text{Được thích}\} \cap \{\text{Được đề xuất}\}}{\{\text{Được đề xuất}\}} = \frac{\text{Số mục được đề xuất chính xác}}{\text{Tổng số mục được đề xuất}} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (6)$$

$$\text{Độ bao phủ } = \frac{\{\text{Được thích}\} \cap \{\text{Được đề xuất}\}}{\{\text{Được thích}\}} = \frac{\text{Số mục được đề xuất chính xác}}{\text{Tổng số mục được thích}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (7)$$

Với những hệ tư vấn được sử dụng để đề xuất các mục dữ liệu cho người sử dụng, đường cong ROC (receiver operating characteristic) sẽ là thích hợp hơn so với đường cong chính xác – bao phủ khi muốn so sánh tính hiệu quả của các hệ tư vấn [10]. ROC dựa trên độ nhạy (sensitivity) và độ đặc hiệu (specificity) và được vẽ theo độ nhạy và phần bù của độ đặc hiệu. Độ nhạy (hay còn được gọi là true positive rate - TPR) có công thức tính như độ bao phủ. Phần bù của độ đặc hiệu (còn được gọi là false positive rate - FPR) được tính theo công thức (8).

$$\text{FPR } = 1 - \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (8)$$

Diện tích dưới đường cong ROC càng lớn thì hiệu suất của giải thuật (hệ tư vấn) càng tốt.

B. Đánh giá hệ tư vấn

Để đánh giá hệ tư vấn, tập dữ liệu (ma trận đánh giá) phải được phân tách thành tập dữ liệu huấn luyện và tập dữ liệu kiểm thử. Tập dữ liệu huấn luyện được sử dụng để học mô hình trong khi tập dữ liệu kiểm thử được sử dụng để kiểm tra mô hình. Các phương pháp phân tách tập dữ liệu có thể là: splitting, bootstrap sampling và k-fold cross validation [15][21]. Phương pháp splitting thực hiện 1 lần phân tách dữ liệu theo tỷ lệ được xác định bởi người sử dụng. Phương pháp bootstrap thực hiện k lần phân tách dữ liệu theo tỷ lệ được xác định bởi người sử dụng. Với phương pháp bootstrap, một người sử dụng có thể là thành viên của tập huấn luyện ở lần phân tách này nhưng là thành viên của tập kiểm kiểm thử ở lần phân tách khác. Phương pháp k-fold cross validation phân tách tập dữ liệu thành k tập con có kích thước bằng nhau và thực hiện k lần đánh giá sau đó lấy kết quả trung bình. Ở mỗi lần đánh giá, (k-1) tập con được sử dụng để học mô hình và 1 tập con còn lại được sử dụng để kiểm tra. Với phương pháp k-fold cross validation, ít nhất một lần, người sử dụng xuất hiện trong tập kiểm thử.

Tập dữ liệu kiểm thử sau đó lại được chia thành hai phần: tập dữ liệu truy vấn (query set) dùng để dự đoán các rating và tập dữ liệu đích (target set) để đánh giá kết quả dự đoán. Việc phân tách tập dữ liệu kiểm thử chỉ được thực hiện khi biết số mục dữ liệu $given$. Với từng người sử dụng, $given$ mục đã được rating sẽ được lựa chọn ngẫu nhiên để đưa cho mô hình hệ tư vấn, các mục đã được rating còn lại được dùng để đánh giá (xem ví dụ ở Bảng 3).

Cách thức đánh giá offline cho một hệ tư vấn lọc cộng tác dựa trên người sử dụng được đề xuất như sau:

Input:
- Ma trận đánh giá;
- Số tập con được sử dụng bởi phương pháp phân tách dữ liệu k-fold cross validation, số mục dữ liệu $given$ để tách tập dữ liệu kiểm thử thành tập dữ liệu truy vấn và tập dữ liệu đích;
- Số láng giềng có độ tương tự cao nhất;
- Số mục dữ liệu cần đề xuất cho người sử dụng.

Output: ma trận nhầm lẫn, độ bao phủ, độ chính xác, độ đặc hiệu, đường cong ROC

Các bước thực hiện:
- Phân tách tập dữ liệu (ma trận đánh giá) thành hai nhóm: tập dữ liệu huấn luyện và tập dữ liệu kiểm thử; sau đó tiếp tục phân tách tập dữ liệu kiểm thử thành: tập dữ liệu truy vấn và tập dữ liệu đích.
- Xây dựng mô hình hệ tư vấn lọc cộng tác dựa trên người sử dụng. Tập dữ liệu huấn luyện được sử dụng để học mô hình hệ tư vấn. Các phép đo Cosine, Pearson, Jaccard và Cohesion được sử dụng để tính sự tương tự giữa những người dùng. Sau đó, với mỗi người dùng trong tập huấn luyện, chọn ra số láng giềng tương tự nhất.
- Đề xuất các mục dữ liệu cho người sử dụng bằng cách dùng mô hình tư vấn ở bước 2 và tập dữ liệu truy vấn.
- Đánh giá tính chính xác giữa tập các đề xuất ở bước 3 với tập dữ liệu đích. Cụ thể, tính độ bao phủ, độ chính xác, độ nhạy, và phần bù của độ đặc hiệu giữa tập các mục dữ liệu được đề xuất với tập dữ liệu đích; sau đó vẽ đường cong ROC.

Bảng 3. Ví dụ về tập dữ liệu truy vấn và tập dữ liệu đích với given=2

Tập dữ liệu kiểm thử									**Tập dữ liệu truy vấn**								
										i1	**i2**	**i3**	**i4**	**i5**	**i6**	**i7**	**i8**
	i1	**i2**	**i3**	**i4**	**i5**	**i6**	**i7**	**i8**	**u2**	3	NA	NA	*NA*	NA	1	NA	NA
u2	3	NA	NA	5	NA	1	NA	NA	**u8**	NA	4	*NA*	NA	*NA*	NA	3	NA
u8	NA	4	3	NA	2	NA	3	NA	**Tập dữ liệu đích**								
									u2	NA	NA	NA	*5*	NA	NA	NA	NA
									u8	NA	NA	*3*	NA	*2*	NA	NA	NA

IV. THỰC NGHIỆM

A. Dữ liệu và công cụ thực nghiệm

MSWeb [1] được sử dụng làm tập dữ liệu thực nghiệm. Dữ liệu của MSWeb được tạo ra bằng cách lấy mẫu và xử lý các nhật ký (log) của www.microsoft.com. Dữ liệu ghi lại việc sử dụng www.microsoft.com của 38000 người dùng ẩn danh được chọn ngẫu nhiên. Với mỗi người sử dụng, dữ liệu liệt kê tất cả các khu vực của website (Vroots) mà người đó truy cập trong khoảng thời gian một tuần của tháng 2 năm 1998. Tập dữ liệu MSWeb dùng trong thực nghiệm chứa 32710 dòng (người dùng hợp lệ), 285 cột (Vroot), 98653 ô chứa giá trị TRUE/1 (các Vroot được truy cập bởi người sử dụng), các ô còn lại (ô thiếu) chứa giá trị FALSE/0 (các Vroot không được truy cập/không được biết bởi người sử dụng).

Bảng 4 hiển thị số lượng người truy cập theo số Vroot. Số lượng người truy cập từ 1 Vroot trở lên (trong 285 Vroot) là 32710 người, từ 2 Vroot trở lên là 22716 (trong 32710 người), từ 10 Vroot trở lên là 875, và từ 19 Vroot trở lên là 45. Số người chỉ truy cập 1 Vroot là gần 10000, chỉ truy cập 2 Vroot là hơn 8000. Vì vậy, để tăng độ tin cậy của các đề xuất của mô hình hệ tư vấn, tập dữ liệu thực nghiệm phải được lọc lại bằng cách thiết lập số lượng tối thiểu (một ngưỡng) các mục dữ liệu được truy cập bởi từng người sử dụng. Trong thực nghiệm này chúng tôi sử dụng ngưỡng là 10.

Bảng 4. Số lượng người truy cập theo Vroot

Số Vroot	>=1	>=2	>=3	>=4	>=5	>=6	>=7	>=8	>=9	>=10	>=11	>=12	>=13	>=16	>=19
Số người truy cập	32710	22716	14283	9544	6280	4151	2767	1871	1281	875	610	432	154	82	45

Tập dữ liệu MSWeb sau khi được chọn lọc lại sẽ được phân tách để xây dựng và đánh giá mô hình tư vấn theo phương pháp k-fold cross validation. Trong thực nghiệm này, số k-fold được chọn là 4. Như vậy, tập dữ liệu được tách thành 4 tập con có kích thước bằng nhau, và mô hình được đánh giá 4 lần sau đó lấy kết quả trung bình. Ở mỗi lần đánh giá, 3 tập con được sử dụng để học mô hình và 1 tập con còn lại được sử dụng để kiểm tra.

Công cụ thực hiện là các hàm do chúng tôi cài đặt kết hợp với một số hàm của gói recommenderlab [15]. Gói recommenderlab là một framework cho phát triển và kiểm thử các giải thuật tư vấn. Các giải thuật được trình bày trong Phần II, III được chúng tôi cài đặt bằng ngôn ngữ R.

B. Kết quả thực nghiệm

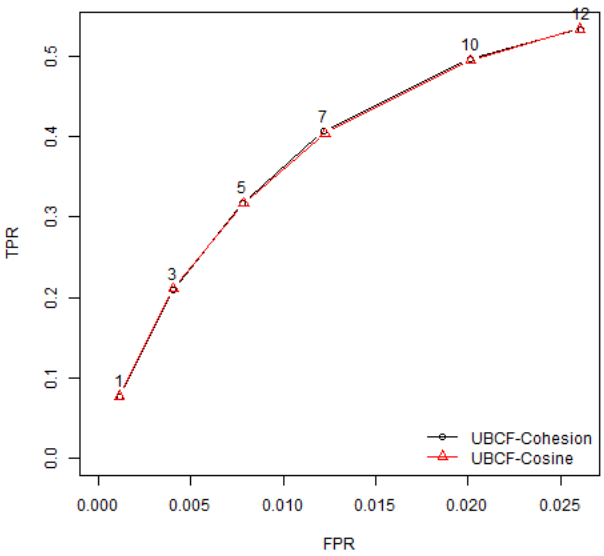
1. Kịch bản 1 – Đánh giá mô hình hệ tư vấn lọc cộng tác dựa trên người sử dụng dùng phép đo Cohesion và Cosine

Các phép đo dùng để đánh giá hệ tư vấn được lưu trong Bảng 5. Bảng này có dòng là số mục dữ liệu được đề xuất cho người sử dụng; cột là các độ đo. Các cột được chia thành 2 nhóm lớn tương ứng với số mục dữ liệu biết trước (given) là 3 và 7. Mỗi nhóm lớn được chia thành 2 nhóm con tương ứng với phép đo Cohesion và Cosine được sử dụng trong hệ tư vấn; mỗi nhóm con lại gồm 3 cột lưu giá trị của độ chính xác, độ bao phủ (độ nhạy) và phần bù của độ đặc hiệu. Nhìn chung, kết quả ở Bảng 5 cho thấy độ chính xác và độ bao phủ của phép đo Cohesion cao hơn của phép đo Cosine. Tuy nhiên, độ chênh lệch là rất thấp.

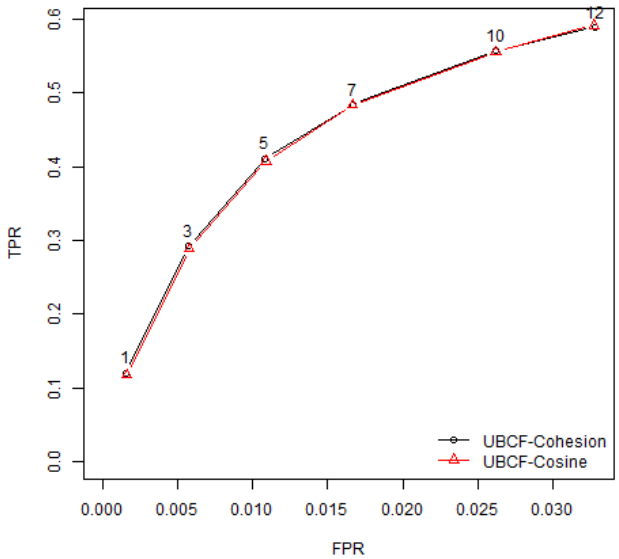
Các đường cong ROC của mô hình hệ tư vấn lọc cộng tác dựa trên người sử dụng dùng phép đo Cohesion và phép đo Cosine tương ứng với số mục biết trước là 3 và 7 được thể hiện trong Hình 2 và 3 tương ứng. Kết quả cho thấy diện tích dưới đường cong ROC của mô hình hệ tư vấn lọc cộng tác dùng phép đo gắn kết Cohesion lớn hơn nhưng sự chênh lệch cũng là rất thấp.

Bảng 5. Bảng lưu giá trị độ chính xác, độ bao phủ (độ nhạy) và phần bù của độ đặc hiệu của 2 phép đo Cohesion và Cosine với given=3 và 7

Số mục đề xuất	Given=3						Given=7					
	Cohesion			Cosine			Cohesion			Cosine		
	Precision	Recall /TPR	FPR	Precision	Recall /TPR	FPR	Precision	Recall /TPR	FPR	Precision	Recall /TPR	FPR
1	0.691176	0.077265	0.001131	0.690045	0.077193	0.001136	0.572398	0.120785	0.001565	0.562217	0.117966	0.001602
3	0.630090	0.210039	0.004066	0.631599	0.210976	0.004050	0.475867	0.292246	0.005757	0.472851	0.289564	0.005790
5	0.574208	0.317800	0.007801	0.571493	0.316625	0.007851	0.409050	0.411220	0.010820	0.406109	0.407469	0.010874
7	0.525372	0.406232	0.012176	0.522140	0.403800	0.012259	0.350194	0.484532	0.016662	0.350194	0.483773	0.016661
10	0.451471	0.495714	0.020105	0.450226	0.494187	0.020151	0.285860	0.557203	0.026168	0.285068	0.555653	0.026197
12	0.406863	0.533379	0.026090	0.407523	0.533631	0.026060	0.253959	0.589483	0.032809	0.255279	0.591863	0.032751



Hình 2. Các đường cong ROC cho lọc công tác dựa trên người sử dụng dùng phép đo Cohesion và Cosine với số mục biết trước là 3



Hình 3. Các đường cong ROC cho lọc công tác dựa trên người sử dụng dùng phép đo Cohesion và Cosine với số mục biết trước là 7

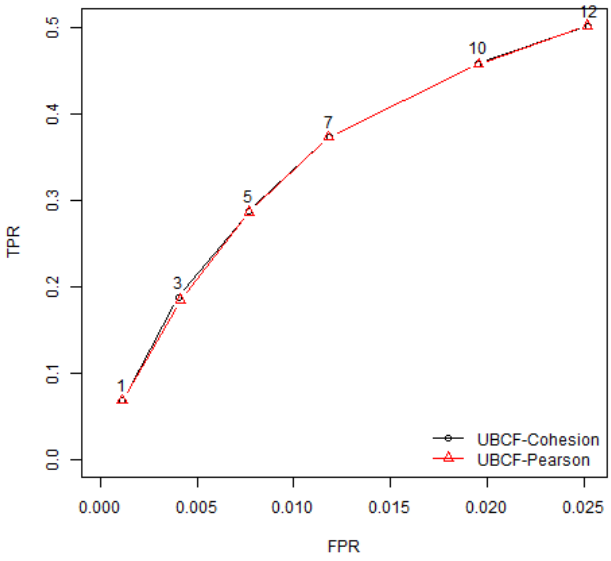
2. Kịch bản 2 – Đánh giá mô hình hệ tư vấn lọc cộng tác dựa trên người sử dụng dùng phép đo Cohesion và Pearson

Các phép đo dùng để đánh giá hệ tư vấn được lưu trong Bảng 6. Bảng này có cấu trúc tương tự như Bảng 5. Nhìn chung, độ chính xác và độ bao phủ của phép đo Cohesion cao hơn của phép đo Pearson. Tuy nhiên, cũng giống như Kịch bản 1, sự chênh lệch là rất thấp.

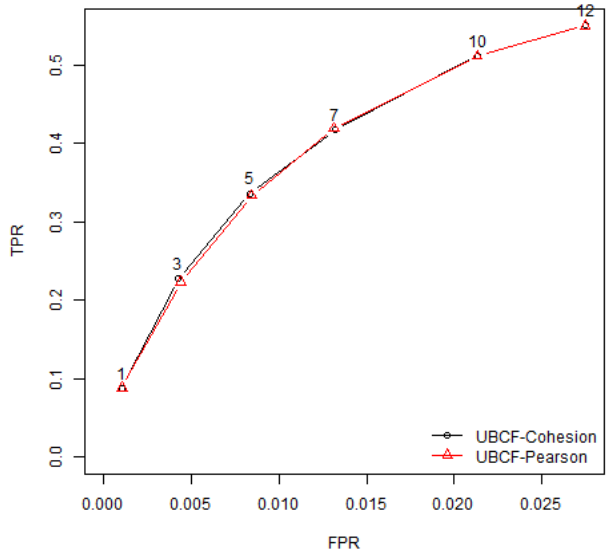
Bảng 6. Bảng lưu giá trị độ chính xác, độ bao phủ (độ nhạy), và phần bù của độ đặc hiệu của 2 phép đo Cohesion và Pearson với given=3 và 5

Số mục đề xuất	Given=3						Given=5					
	Cohesion			Pearson			Cohesion			Pearson		
	Precision	Recall /TPR	FPR	Precision	Recall /TPR	FPR	Precision	Recall /TPR	FPR	Precision	Recall /TPR	FPR
1	0.694805	0.069056	0.001123	0.698052	0.069307	0.001110	0.707792	0.087818	0.001073	0.712662	0.088249	0.001055
3	0.634199	0.187631	0.004035	0.625000	0.185270	0.004138	0.611472	0.227389	0.004285	0.600108	0.223417	0.004411
5	0.582468	0.286944	0.007679	0.581818	0.286763	0.007691	0.545455	0.336459	0.008357	0.540584	0.333871	0.008448
7	0.540816	0.372704	0.011826	0.541976	0.373680	0.011796	0.487245	0.417316	0.013200	0.489796	0.419539	0.013135
10	0.468994	0.458266	0.019537	0.468344	0.457584	0.019561	0.420942	0.512364	0.021301	0.420455	0.511613	0.021320
12	0.428707	0.501477	0.025227	0.429654	0.502161	0.025184	0.378111	0.550778	0.027457	0.377976	0.550585	0.027464

Các đường cong ROC của mô hình hệ tư vấn lọc cộng tác dựa trên người sử dụng dùng phép đo Cohesion và phép đo Cosine tương ứng với số mục biết trước là 3 và 5 được thể hiện trong Hình 4 và 5 tương ứng. Kết quả cho thấy diện tích dưới đường cong ROC của mô hình hệ tư vấn lọc công tác dùng phép đo gắn kết Cohesion lớn hơn nhưng sự chênh lệch cũng là rất thấp.



Hình 4. Các đường cong ROC cho lọc công tác dựa trên người sử dụng dùng phép đo Cohesion và Pearson với số mục biết trước là 3



Hình 5. Các đường cong ROC cho lọc công tác dựa trên người sử dụng dùng phép đo Cohesion và Pearson với số mục biết trước là 5

3. Kịch bản 3 – Đánh giá mô hình hệ tư vấn lọc cộng tác dựa trên người sử dụng dùng phép đo Cohesion và Jaccard

Tương tự như hai kịch bản trước, các phép đo và đường cong ROC dùng để đánh giá hệ tư vấn lọc cộng tác dựa trên người sử dụng dùng phép đo Cohesion và Jaccard được lưu trong Bảng 7 và các Hình 6, 7 tương ứng.

Bảng 7. Bảng lưu giá trị độ chính xác, độ bao phủ (độ nhạy), và phần bù của độ đặc hiệu của 2 phép đo Cohesion và Jaccard với given=5 và 9

Số mục đề xuất	Given=5						Given=9					
	Cohesion			Jaccard			Cohesion			Jaccard		
	Precision	Recall /TPR	FPR	Precision	Recall /TPR	FPR	Precision	Recall /TPR	FPR	Precision	Recall /TPR	FPR
1	0.644796	0.093813	0.001301	0.644796	0.094062	0.001301	0.446833	0.168027	0.002023	0.429864	0.159841	0.002086
3	0.567119	0.245593	0.004757	0.567873	0.246112	0.004749	0.329563	0.339529	0.007361	0.327677	0.329218	0.007381
5	0.503620	0.359253	0.009091	0.500000	0.356187	0.009158	0.264932	0.438207	0.013457	0.265385	0.438198	0.013448
7	0.448933	0.444474	0.014132	0.445055	0.440562	0.014231	0.227214	0.518517	0.019812	0.226729	0.513673	0.019824
10	0.372964	0.520996	0.022976	0.370814	0.518873	0.023057	0.184163	0.586594	0.029890	0.183258	0.582452	0.029923
12	0.335784	0.561837	0.029212	0.336538	0.563204	0.029179	0.163650	0.618518	0.036778	0.162990	0.614595	0.036806

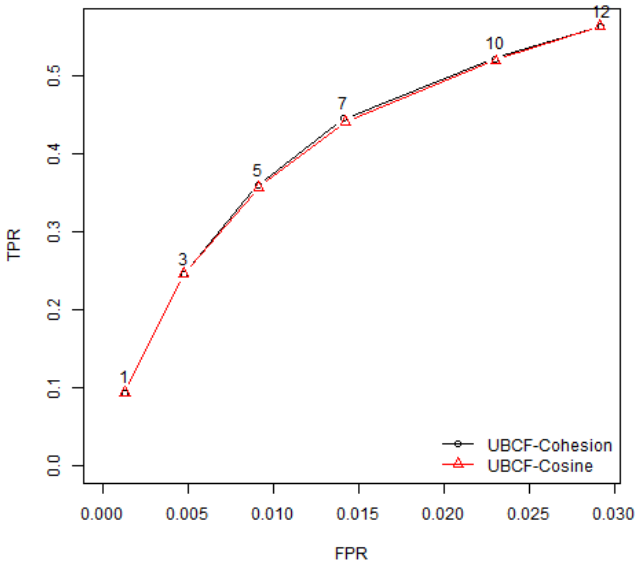
Kết quả trong các bảng và hình cho thấy độ chính xác và độ bao phủ của phép đo Cohesion cao hơn của phép đo Jaccard; diện tích dưới đường cong ROC của mô hình hệ tư vấn lọc công tác dùng phép đo gắn kết Cohesion lớn hơn diện tích dưới đường cong ROC của mô hình hệ tư vấn lọc công tác dùng phép đo gắn kết Jaccard. Tuy nhiên, cũng giống như kết quả ở hai kịch bản trước, sự chênh lệch giữa tính chính xác của hai mô hình là rất thấp.

4. Đánh giá chung

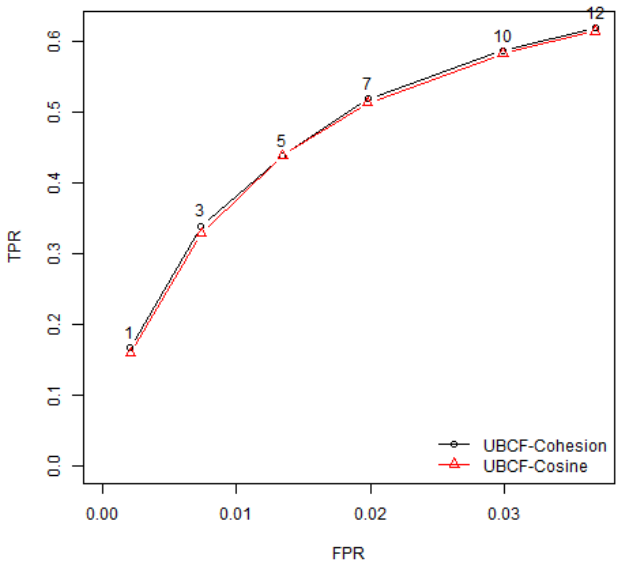
Trong hệ tư vấn lọc cộng tác dựa trên người sử dụng, các phép đo Cosine và Pearson là hai phép đo phổ biến để tính sự tương tự giữa người sử dụng. Nếu tập dữ liệu ở dạng 0 – 1 thì phép đo Jaccard cũng thường được sử dụng để tính sự tương đồng. Tuy nhiên, ba phép đo này đều ở dạng đối xứng.

Với phép đo gắn kết hàm ý thống kê Cohesion – một phép đo không đối xứng, kết quả đánh giá mô hình hệ tư vấn lọc cộng tác dựa trên người sử dụng dùng phép đo này cho thấy nó có độ chính xác cao hơn so với các mô hình dùng phép đo Cosine, Pearson và Jaccard. Tuy nhiên, sự chênh lệch là không nhiều.

Mặc dù mô hình hệ tư vấn dùng phép đo Cohesion có độ chính xác cao hơn nhưng việc tính giá trị tương tự giữa m người sử dụng theo phép đo Cohesion là lâu hơn so với 3 phép đo còn lại. Nguyên nhân là phép đo Cohesion phải tính giá trị gắn kết (u_a, u_b) và giá trị gắn kết (u_b, u_a) trong khi phép đo Cosine, Peason và Jaccard chỉ cần tính một lần do đặc điểm đối xứng. Vì vậy, tùy vào bài toán thực tế, nếu cần phân biệt chiều tác động giữa hai người sử dụng, phép đo Cohesion nên được sử dụng; ngược lại ta có thể sử dụng một trong 3 phép đo kia để rút ngắn thời gian thực thi.



Hình 6. Các đường cong ROC cho lọc công tác dựa trên người sử dụng dùng phép đo Cohesion và Jaccard với số mục biết trước là 5



Hình 7. Các đường cong ROC cho lọc công tác dựa trên người sử dụng dùng phép đo Cohesion và Jaccard với số mục biết trước là 9

V. KẾT LUẬN

Bài báo đề xuất một cách tiếp cận mới trong tư vấn lọc cộng tác dựa trên người sử dụng qua việc dùng phép đo gắn kết hàm ý thống kê Cohesion để tính sự tương tự giữa hai người sử dụng. Phép đo Cohesion là không đối xứng nên phù hợp với thực tế vì sự ảnh hưởng qua lại giữa hai người sử dụng là không đối xứng. Mô hình tư vấn lọc cộng tác theo đề xuất đã được đánh giá và so sánh với mô hình tư vấn lọc cộng tác theo 3 phép đo tương tự phổ biến nhất Cosine, Pearson và Jaccard. Trên tập dữ liệu thực nghiệm MSWeb, kết quả thực nghiệm cho thấy mô hình tư vấn sử dụng phép đo Cohesion nhìn chung có ưu thế hơn so với các mô hình tư vấn còn lại nhưng sự chênh lệch là không nhiều.

VI. LỜI CẢM ƠN

Nghiên cứu này do Dự án UK – ASEAN Research Hub, Viện Nghiên cứu và Đào tạo Việt - Anh, Đại học Đà Nẵng tài trợ (07/HĐ-UARH).

TÀI LIỆU THAM KHẢO

[1] A. Asuncion, D. J. Newman, UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science.http://www.ics.uci.edu/~mlearn/MLRepository.html, 2007.

[2] A. Felfernig, M. Jeran, G. Ninaus, F. Reinfrank, S. Reiterer, and M. Stettinger, “Basic Approaches in Recommendation Systems”, Recommendation Systems in Software Engineering, pp. 15-38, Springer-Verlag Berlin Heidelberg, 2014.

[3] A. Said, D. Tikk, A. Hotho, “The Challenge of Recommender Systems Challenges (tutorial)”, ACM RecSys’12 - Proceedings of the sixth ACM conference on Recommender systems, pp.1-2, 2012.

[4] A.N. Nikolakopoulos, M. A. Kouneli, and J. D. Garofalakis, “Hierarchical Itemspace Rank: Exploiting hierarchy to alleviate sparsity in ranking-based recommendation”, Neurocomputing, Volume 163, pp. 126-136, Elsevier, September 2015.

[5] C. Desrosiers, and G. Karypis, “A Comprehensive Survey of Neighborhood-based Recommendation Methods”, Recommender Systems Handbook, pp. 107-144, Springer US, 2011.

[6] D. M. Ekstrand, J. T. Riedl, J. A. Konstan, Collaborative Filtering Recommender Systems, Journal Foundations and Trends in Human-Computer Interaction Volume 4 Issue 2, February 2011, Pages 81-173.

[7] F. O. Isinkaye, Y. O. Folojimi, B. A. Ojokoh. Recommendation Systems: Principles, Methods an evaluation, Egyptian Informatics Journal, Volume 16, Issue 3, pp: 261–273, 2015.

[8] F. Ricci, L. Rokach, B. Shapira, P.B. Kantor, Recommender Systems Handbook, Springer US, 2011.

[9] G. Adomavicius and A. Tuzhilin, “Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions”, IEEE Transaction Knowledge and Data Engeering, vol. 17, no. 6, pp. 734-749, June 2005.

[10] G. Shani and A. Gunawardana, “Evaluating recommendation systems”, Recommender Systems Handbook, pp. 257–97, Springer US, 2011.

[11] H. Steck, “Evaluation of Recommendations: Rating-Prediction and Ranking”, In Proceedings of the 7th ACM conference on Recommender systems (RecSys '13). ACM, New York, NY, USA, 493-494, 2013.

[12] I. Avazpour, T. Pitakrat, L. Grunske, and J. Grundy, “Dimensions and Metrics for Evaluating Recommendation Systems”, Recommendation Systems in Software Engineering, pp. 245-274, Springer-Verlag Berlin Heidelberg, 2014.

[13] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative Filtering Recommender Systems", The Adaptive Web, LNCS 4321, pp. 291-324, Springer-Verlag Berlin Heidelberg, 2007.

[14] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering", Proceedings of the 14th Conf. Uncertainty in Artificial Intelligence, July 1998.

[15] M. Hahsler, "Recommenderlab: A Framework for Developing and Testing Recommendation Algorithms", Southern Methodist University, 2011

[16] M. Millan, M.F. Trujillo, E.Ortiz, A Collaborative Recommender System Based on Asymmetric User Similarity. IDEAL 2007, LNCS 4881, Springer, 2007, pp. 663-672.

[17] M. J. Pazzani1 and D. Billsus, "Content-Based Recommendation Systems", The Adaptive Web, LNCS 4321, pp. 325-341, Springer-Verlag Berlin Heidelberg, 2007.

[18] P. Pirasteh, D. Hwang, J. J. Jung, "Exploiting matrix factorization to asymmetric user similarities in recommendation systems", Journal Knowledge-Based Systems Volume 83 Issue C, pp. 51-57, Elsevier Science Publishers, July 2015.

[19] R. Gras and P. Kuntz, "An overview of the Statistical Implicative Analysis (SIA) development", Statistical Implicative Analysis - Studies in Computational Intelligence (Vol. 127), pp.11-40, Springer-Verlag, Berlin-Heidelberg, 2008.

[20] R. Gras et al., L'implication statistique – Nouvelle méthode exploratoire de données, La pensée sauvage édition, 1996.

[21] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection." In Proceedings of the Fourteenth International Joint Conference on Arti-ficial Intelligence, pp. 1137–1143, 1995.

[22] Y. Koren, and R. Bell, "Advances in Collaborative Filtering", Recommender Systems Handbook, pp. 145-186, Springer US, 2011.

[23] Z. L. Zhao, C. D. Wang, and J. H. Lai, AUI&GIV: Recommendation with Asymmetric User Influence and Global Importance Value, PLOS ONE | DOI:10.1371/journal.pone.0147944 , February 1, 2016

USER-BASED COLLABORATIVE FILTERING RECOMMENDATION USING THE STATISTICAL IMPLICATIVE COHESION MEASURE

Lan Phuong Phan, Trang Uyen Tran, Hung Huu Huynh, Hiep Xuan Huynh

ABSTRACT—*Similarity measures play an important role in collaborative filtering recommender systems. Pearson correlation coefficient, Cosine similarity, and Jaccard coefficient are the most common similarity measures used for calculating the similarity between users (or items). These measures are the symmetric while in fact, the interaction between users is often asymmetric. This paper proposes a new approach in collaborative filtering method to calculate the similarity of each pair of users by using the statistical implicative cohesion measure Cohesion, as well as focuses on evaluating recommender systems that use different similarity measures. The k-fold cross validation method and the MSWeb dataset are used for the experiment. The result shows that collaborative filtering recommender system using Cohesion is better than collaborative filtering recommender systems using Cosine, Pearson, and Jaccard.*

Keywords*— User-based collaborative filtering, statistical implicative cohesion measure, recommender system.*